Measuring the Impacts of Security Interventions: An Introduction

Franziska Seethaler UNU/SIPA Junior Research Fellow

Executive Summary

Assessing the impact of security interventions is a challenging – yet essential – endeavour. Without adequate expertise, resources, and political will, impact assessments are unlikely to contribute effectively to rule of law programming, drug control policy, efforts to tackle organized crime, or countering violent extremism (CVE) interventions. Rigorous impact assessment, on the other hand, can stimulate innovation, and improve both the effectiveness and efficiency of security interventions – important for beneficiary communities and financially-strapped donors alike. However, effective impact assessment in volatile environments is challenging: insecurity itself complicates assessment design, implementation, and analysis.

This paper provides an introduction to the challenges of effective security intervention impact assessment. It discusses different measurement tools and methods available to security practitioners and programmers, and offers ideas for strengthening impact assessment in these fields.

Impact assessment is often not prioritized when interventions are conceptualized and planned. It is only tacked on once accountability to donors needs to be demonstrated. This makes programmatic progress difficult to measure, because no thorough baseline studies are conducted when the programmes are set up. Security practitioners are often not conversant with assessment tools and methods, measurement design, data collection, and analysis. Instead of undertaking the hard task of measuring long-term impacts, programmers often resort to measuring easily quantifiable outputs, which fail to provide insights into an intervention's complex impact. This risks wasting scarce donor money, producing poorly-tailored interventions – or worse, doing real, but unrecognized, harm to supposed beneficiaries.

Many of the challenges discussed in this paper can be mitigated through thoughtful application and combination of different methods and tools. The second part of the paper therefore provides an introduction to basic assessment methods including surveys, experiments, and select qualitative approaches, explaining how they work, their potential advantages and drawbacks, and offering examples of their use in relation to security interventions.

The final part of the paper highlights new possibilities for more accurate, comprehensive, and thoughtful approaches to impact assessment of contemporary security interventions. These include the early integration of impact assessments into programme design, lengthened time horizons, as well as the use of mixed methods, experimental designs, and new technologies.

The paper also points out, however, that reaping the full benefits of the methodological innovations requires a corresponding cultural change, generating greater familiarity amongst practitioners and donors with different approaches to – and benefits from – impact assessment. This should be matched by a recalibration of expectations, time frames, and budgets for impact assessment.

Contact: Franziska Seethaler, fms2143@columbia.edu

1. INTRODUCTION

How can we measure the impact of security interventions[1] in the areas of the rule of law, drug control policy, organized crime, and countering violent extremism (CVE)? Understanding the impact of these security interventions allows for the modification of future project design, tailored programming, improved efficacy, and targeted allocation of scarce resources; all important to both recipient communities and fiscally-strained donors alike.

Accurately assessing the complex impact of contemporary security interventions is inherently challenging. Choosing a well-suited assessment methodology - essential for making inferences about causal relationships - requires expertise, institutional capacity, financial resources, access, and political support. Methodological choices, data collection, and analysis are further complicated by the nature of the volatile conflict environments in which security interventions occur. The aforementioned components of rigorous impact assessments are thus often lacking. Instead impact assessments are conducted with a narrow focus on easily accessible and quantifiable metrics, such as the number of weapons collected, drugs seized, or combatants cantoned. These metrics, however, are not necessarily accurate measures of an intervention's complex impact. As a consequence, ill-conceived impact assessments can on occasion provide wildly inaccurate impressions of an intervention's efficacy thus wasting scarce donor money. They may also have unintended negative consequences for the target population as causation is misattributed and programmes are not modified accordingly.

Despite or perhaps because of these challenges involved in impact assessment, there is a growing debate about how best to assess intervention impacts. This intensifying debate has already produced some more thoughtful and tailored approaches to impact assessment, encouraging a longer-term perspective on programming and generating support for initiatives to identify lessons learned and to facilitate cross-pollination across fields. Methodological and technological innovations, such as experimental assessment designs and the use of new mobile technologies in data collection and analysis, also offer new opportunities. While these innovations are not a panacea to continued problems, which will require additional attention and innovation, they put more accurate and cost-effective measurement within reach thus enhancing our capacity to design more effective interventions.

Recently, the importance of rigorous measurement has also been emphasized in the context of the Sustainable Development Goals (SDGs) that were adopted at the UN Sustainable Development Summit on 25-27 September 2015.[2] The SDGs reflect an expanded understanding of sustainable development, beyond that of the Millennium Development Goals (MDGs). SDG 16 for example calls for the reduction of "all forms of violence", as well as "efforts to combat organized crime and to promote the rule of law."[3] With regard to measurement, the Transforming our World: 2030 Agenda for Sustainable Development highlights the need to strengthen data collection and capacity building for broader, high-quality measurement of progress toward the SDGs.[4]

In light of these developments, this working paper seeks to provide an overview of the current state of impact assessment in security interventions and explore the various assessment methods at programmers' disposal. The first part of this working paper focuses on current challenges of impact assessment and highlights evolving approaches of impact assessment and Monitoring and Evaluation (M&E) more generally. The second part provides an introduction to basic assessment methods, areas of application, and respective advantages and drawbacks. New possibilities for more accurate, comprehensive, and thoughtful approaches to M&E will be highlighted in the last part of this paper. An appendix provides a tabulated bibliography, offering leads for further reading on how the different methods and tools discussed in this paper may be applied in various areas of security interventions.

2. THE CHALLENGE OF ASSESSING SECURITY INTERVENTION IMPACTS

Rigorous impact assessment is inherently difficult, requiring sufficient expertise, resources, access, and political will to include M&E from the conceptualization phase of security interventions. Impact assessment is particularly difficult to conduct in the volatile environments of complex contemporary security interventions because gaining the necessary access and collecting the data is challenging. There is growing recognition that traditional approaches to impact assessment in security interventions have been flawed: many programmatic assessments were limited to a rudimentary evaluation carried out over a short time span, and focused on assessing easily measurable outputs such as the number of arrests in organized crime interventions or the number of drugs seized in drug policy interventions.[5] Rarely were impact assessments oriented to effectively measure the larger effects of programming or to provide sufficient context for analysing impact.[6]

The consequences of continuing to conduct these types of assessments are laid bare by the example of Mozambique's disarmament, demobilization, and reintegration (DDR) programme. An assessment of the United Nations Mission in Mozambique (ONUMOZ) DDR programme focused on the number of weapons it collected as part of its disarmament mandate. Between 1992 and 1994, ONUMOZ collected 200,000 weapons,[7] which in absolute terms appeared to represent a success, but in context, this number represented only a small fraction of the millions of weapons estimated to be in Mozambique at the time.[8] Moreover, the programme failed to put the weapons it had collected beyond use – and

the caches were eventually raided by criminal elements. Mozambican arms were then smuggled to surrounding states, where they fuelled further conflict and crime.[9]

Although there is intensifying debate surrounding, and growing recognition of, the need to improve the rigor of impact assessment and to better tailor M&E exercises to realities on the ground, it is unclear if there is yet widespread understanding and application of innovative methods and technologies amongst programme staff and donors. Persisting challenges with regard to assessment design, implementation, and analysis will therefore be outlined in the following section.

a) Assessment Design

As the Mozambican example illustrates, a narrow focus on short-term programmatic outputs, instead of the long-term impacts of security interventions, can overlook the complex effects of these programmes and misconstrue the security intervention's efficacy. In the case of Mozambique, a focus on the number of weapons collected failed to give programmers an indication of the total number of weapons circulating in Mozambique, nor an understanding of recidivism rates or the dynamics of informal criminal elements. A focus on outputs rather than outcomes (or impacts) can therefore ultimately prevent the necessary modification of programmes, which in turn can negatively affect the target population and undermine their trust in the programmes.

The issue of output- versus impact-related indicators is not only relevant in the context of DDR programmes. It is also reflected in the current debate around drug policy. Instead of merely focusing on immediate enforcement outputs, such as asset seizures or arrests, there has been increased interest in measuring the larger societal impact of organized crime policy (e.g. long-term reduction of violence and vulnerabilities).[10] Growing criticism queries whether an approach that focuses only on outputs will adequately reflect progress (or lack thereof) towards the larger goals highlighted by the international drug control conventions, such as improving the health and welfare of populations.[11] In addition to criticism of output-oriented metrics, there has been increasing scepticism regarding approaches that fail to take into account knock-on effects (e.g. eradication in one area stimulating production in another) and approaches that use inadequate metrics, in particular in cases where direct measurement is difficult. Situations where observation is problematic and direct measurement is not possible pose additional challenges, requiring alternate methods and adequate proxy indicators that allow for indirect measurement.

Assessing the impact of preventive programmes represents another challenge. For example, how can programmers demonstrate that their interventions aimed at preventing the proliferation of organized crime work? If no new incidents of organized crime occur in an area, is this phenomenon attributable to the preventive programmes or to other factors in the programme environment? One option to attempt to measure the impact of these programmes is counterfactual analysis. Both scenarios with and without the intervention are compared, in order to establish cause and effect between the intervention and the observed outcomes. However, as illustrated with the organized crime example above, even in cases where a change in conditions is observable, it might be more difficult to determine whether this change was unequivocally induced by the intervention itself or whether conditions improved due to other external factors (the 'attribution problem').

In recent years, the United Nations and other actors have increasingly recognized that we need to not only rethink what we seek to measure, but how we seek to measure it. There is increasing recognition that better metrics and measurement alone are not the answer, but rather a more holistic approach is needed in order to adequately reflect the complex impact of interventions on realities on the ground. This does not only include the aforementioned issue of output- versus impact-related indicators, but also concerns about the level of analysis, metrics used, and sample size. Instead of – or in addition to – measuring the number of court cases processed by a newly reformed judiciary, for example, measuring the public's acceptance of, and experience with, the new institutions may be a more effective approach to understanding the impact of such reforms.[12] As the choice of methods and metrics is subjective and context-dependent, the advantages and shortfalls of this choice need to be discussed in the analytical stage of impact assessment.

b) Assessment Implementation

One of the main obstacles to effective measurement is whether the data needed can be accessed in the time frame and with the resources available. This includes issues of access, but also security constraints in conflict-ridden contexts where security interventions take place. In these contexts, access to baseline data can pose a problem, for example in cases where the interventions were set up under time pressure and a thorough collection of pre-intervention baseline data was not feasible. This lack of baseline data can distort the impact assessment because no data is available against which to measure progress.

Both ethical concerns and privacy issues can also limit certain types of measurement. There have been instances where ethical concerns have been voiced with regard to programmes that include a random assignment to treatment and control groups, the latter of which are "denied" treatment. In other cases, privacy concerns have arisen with regard to the personal data compiled for intervention assessment, its storage, and the potential it might be shared with outside actors. For example, in the case of some DDR programmes, biometric data – such as fingerprints – are collected during the registration of ex-combatants and

personal data including information about membership in armed groups might be stored in databases. Privacy concerns about this information have the potential to impede data access in impact assessment, both during data collection and in subsequent inter-agency information sharing. If not for strong safeguards, privately- or nationally-run DDR programmes might have reservations to share personal data with other national and international actors. All of these constraints can result in a small or unrepresentative sample size that is not sufficiently large to produce reliable results. Maintaining high research and ethical standards in these cases is of utmost importance for the sustainability and success of the programmes.

c) Assessment Analysis

In the iterative process of programme design, quality analysis is needed to match the data collection. In this regard, challenges may arise in the synthesis, analysis, and contextualization of data, as well as the presentation of results. The complexity of conflict dynamics and the interconnectivity of different factors can complicate the identification of causal relationships and thus make it difficult to attribute specific impacts to particular security interventions. In the field of organized crime interventions, the Central American Regional Security Initiative (CARSI), an integrated rule of law programme partly funded by the U.S. Department of State, provides a pertinent example: While researchers were able to demonstrate a reduction of municipal-level crime through programmes for at-risk youth and communities, they were unable to attribute the success to a specific programme as disparate initiatives (e.g. counter-gang, neighbourhood clean-up, improved lighting programmes) were launched in concert.[13] In addition to these attribution problems, biases, spill-over effects, and interdependent variables can distort the findings. For example, it might be difficult to discern whether an intervention or another factor, such as the growing economy, was responsible for the improved security situation in a country. Depending on how data are analysed, the same data can yield very different results and thus be left open to alternate, even conflicting, interpretations.

The challenges and the benefits of more rigorous approaches to impact assessment are not limited to one area of security interventions. As such, advances in one area can facilitate learning across security intervention fields. Innovations in one field – such as anti-gang programming in North America, which has through innovation and analysis developed sophisticated measurement techniques over the past two decades – may also be highly informative for programme design, implementation, and data analysis in other fields, such as CVE.[14]

Numerous challenges discussed in this paper can be mitigated through the thoughtful application and skilful combination of the methods and tools available. For that reason, the following sections seek to provide an introduction to the measurement methods and tools available to practitioners working in the areas of drug control policy, organized crime, rule of law, and CVE. The following chapters do not have the aspiration of being exhaustive,[15] but rather focus on the most pertinent methodological approaches and metrics available to security intervention programmers: surveys, experiments, indicators, indices, benchmarks, as well as other selected qualitative approaches. To increase readability, each section will provide a description of the method, explain the main variants or types, highlight areas of application, and discuss its potential advantages and drawbacks.

3. SURVEYS

a) Description

Survey research encompasses any measurement procedures that involve asking questions of respondents, ranging from in-depth in-person interviews to short, automated telephone questionnaires. Depending on the scope and breadth of the inquiry, surveys can take either qualitative (e.g. in-depth interviews of a small number of people) or quantitative form (e.g. questionnaires distributed to a large sample of a population). Surveys, when conducted on a larger scale, allow researchers to measure and make statistical inferences about the attitudes and behaviours of a large sample of participants.[16]

b) Types

- **Questionnaires:** Questionnaires or social surveys are a method used to collect standardized data in a statistical form from a large number of people. Types of questionnaires include factual surveys, which collect descriptive information (e.g. government census), attitude surveys, and explanatory surveys.[17] Questionnaires can be conducted in person or via mail, email, online, or by telephone

- **Interviews:** Face-to-face interviews can be conducted individually or in focus groups. Individual interviews help foster an atmosphere of trust and privacy in order to gain a deeper understanding of the interviewees' attitudes, beliefs, and experiences. The purpose of focus group interviews is to learn through discussion in small groups about "conscious, semiconscious, and unconscious psychological and socio-cultural characteristics and processes among various groups".[18]

Survey questions can be standardized (to enhance comparability), unstandardized (useful in pilot studies when researchers are still formulating survey questions) or semi-standardized (using a set of pre-determined questions as a guideline, but leaving room for flexibility).[19] Open-ended or closed-ended questions can be utilized depending on the context, expected use, and goals. Instead of using simple yes/no answers, scaled responses allow respondents to give more nuanced answers. The so-called Lickert Scale uses a scale

from 1 to 5 to systematize respondents' answers, which also allows for easier graphical analysis of responses. The Hague Institute for the Internationalization of Law, for example, utilized the Lickert Scale in their assessment of justice needs and access to justice in Mali by applying a scale from 1 to 5 to systematize answers to the 110 questions they asked to survey participants.[20]

c) Areas of Application

Impact assessments of security interventions frequently include surveys. In the area of disarmament, demobilization, and reintegration, for example, Christopher Blattman and Jeannie Annan conducted a randomized survey-based assessment to analyse the incentives of child involvement in armed groups.[21] Macartan Humphreys and Jeremy Weinstein conducted a survey of 1,043 ex-combatants in Sierra Leone in 2003, to analyse which determinants facilitate political, social, and economic reintegration of ex-combatants.[22] In research on CVE, Vanessa Corlazzoli used a survey to establish a baseline assessment for the Countering and Preventing Radicalization project in Indonesia, inquiring about the availability of places of worship and the community's tolerance of different religions and attitudes towards women.[23] In their assessment of the justice sector in Yemen, Martin Gramatikov et al. used survey data to measure the public's acceptance and experience with justice institutions.[24]

d) Advantages

Surveys are a tool that allows researchers to understand the subjective attitudes of a large number of people and to monitor changes in perception. When designed and executed properly, surveys enable researchers to make inferences about larger populations. In addition to proper sampling, tailoring the survey for particular audiences enhances the quality of this method. To reach respondents of different backgrounds, it may be necessary to adapt the order, content, and style of the survey questions to the respective audience. In case of illiterate respondents, for example, verbal interviews may be employed. Depending on the context, standardized, semi-standardized, and unstandardized surveys can be employed that provide the researcher with the necessary flexibility to either adapt questions to the respective audience and/or enhance comparability across a larger group of respondents. Surveys can facilitate the collection of a great depth or breadth of data.[25] When conducted using information and communication technologies, instead of enumerated interviews, surveys can be administered remotely, expanding the breadth of the study.[26] Software can also make it possible to analyse large amounts of data quickly at low cost, using sophisticated statistical techniques.

e) Drawbacks

The effectiveness of survey research is diminished when issues of poor survey design, implementation, and analysis are not adequately addressed. With regard to survey design, the choice of the type of survey used should take into consideration the demographics of the population, cultural specificity, and the degree of flexibility needed to obtain the answer to the questions the enumerator seeks to answer.[27] For example, if researchers administer a written questionnaire in regions with low literacy rates, they may introduce a major bias into the study. Furthermore, survey questions must be carefully phrased in order to prevent respondents resorting to politically correct answers to avoid presenting themselves in an unfavourable manner or out of fear of negative repercussions,[28] when presented with questions about controversial issues.[29] In addition, double-barrelled, biased, or politically sensitive questions might confuse or irritate the respondents, thus distorting the data.[30] Another topic that needs to be addressed is the issue of human subjects protections. By demonstrating to participants that their answers are confidential and the data is protected, they may feel more at ease and are able to answer more honestly. The use of strong ethical safeguards, such as human subjects review and approval, and security safeguards in the handling of data provided by respondents can add to the complexity, cost, and timeframe of administering surveys effectively, but also improves the reliability and robustness of the evidence thus gathered.

When implementing surveys, the possibility of biases, such as sampling effects that distort findings must also be considered.[31] Conducting a successful survey also depends on accessing the target population, which may be especially difficult in conflict contexts or other environments with security risks or concerns. Access issues need to be addressed through the choice of research design and data collection methods, including mobile technologies. One of the main challenges of survey analysis is the interpretation of non-responses. Non-responses, especially in online and mobile surveys, can lead to misperception by surveyors of the characteristics or attitudes of the population for which they are designing the intervention.[32] Consequently, the challenges of interpreting non-responses as well as survey answer options, such as the meaning of options like "partially agree" for example, needs to be addressed in the analysis of the survey data collected.

Aside from inherent limitations to certain survey approaches, most of the problems lie in poor survey design, implementation, and analysis. Yet through a solid survey design and a rigorous analysis, the drawbacks of survey design can be reduced.

4. EXPERIMENTS

a) Description

Experimental designs, commonly used in the natural sciences, allow researchers to make claims about causal inference by systematically controlling one or more independent vari-

ables (i.e. causal factors) and measuring any change in the dependent variable (i.e. effect).[33] When applied to security interventions, experiments seek to determine the effects of a certain type of treatment (e.g. security intervention such as the presence of peacekeepers) on a target population. In order to establish that the security intervention is indeed responsible for the observed outcome, *control groups* are frequently employed.[34] In comparison to the *treatment group,* the *control group* – often unknowingly – receives no exposure to security intervention (e.g. access to DDR programming). Sometimes researchers make use of the so-called "placebo effect" to test if the same effects occur when an inert or non-treatment is administered, simply due to the recipients' expectation that they are part of the treatment group

b) Types and Characteristics

Myriad types of experiments exist, of which only the most frequently used in security interventions will be highlighted in the following section. The types and characteristics in this list are not mutually exclusive, but some characteristics, such as double-blind experiments and matched subject designs, can be used in combination.

- **Within-group and between-group designs:** To compare the impact of different treatments, within-group designs expose every participant to a treatment and then measure its effect. In the second step, they expose all participants to another treatment and measure the latter's effect, an approach that requires fewer participants and mitigates the lack of a control group. In cases where a control group is available, participants in between-group designs are assigned to either treatment or control group, in order to measure the effects of systematically controlling one or more causal factors.[35]
- **Randomized controlled trials (RCTs):** In RCTs, participants are randomly assigned to the treatment or the control group, allowing for the effectiveness of the treatment to be compared against the baseline that the control group establishes.[36] Randomization in RCTs helps reduce sampling biases of the researcher.
- **Quasi-experimental designs:** Like in experimental designs, quasi-experimental designs divide participants into treatment and control groups in order to test claims about causal inferences, but lack the randomized assignment used in RCTs. They are frequently used in individual case studies or in situations where randomization is difficult.[37]
- **Stepped-wedge designs:** Stepped-wedge designs involve a sequential rollout of an intervention to participants. Due to the staggered intervention over multiple measurement periods, no division into treatment and control group is needed, as participants with later access to the intervention serve as the control group with which to compare the effects of the programme.

Consequently, a smaller sample is needed in stepped-wedge designs. Stepped-wedge designs also have logistical and financial advantages, as interventions are seldom carried out simultaneously for all participants. Stepped-wedge designs can also help to minimize ethical issues as participants are not "denied" treatment, but advantage is taken of inherent rollout delays to test the efficacy of programmes.
- **Matched subject designs:** When RCTs or large-scale sampling is problematic, matched subject designs attempt to enhance the comparability of small control and treatment groups, by matching subjects across groups based on pre-existing characteristics, such as race or gender.
- **Double-blind experiments:** In double-blind experiments neither the participants, nor the researcher know whether the participants have been assigned to the treatment or control group. Consequently, double-blind experiments can help to reduce biases.
- **Counterbalanced measures designs:** In cases where having a control group is problematic, a counterbalanced measures design allows all participants to have the treatments, but alters the order in which they receive them to test the impact of the order of the treatment.[38]
- **Behavioural games:** Behavioural games use experimental settings to make inferences about beliefs and behaviours. The objective is to analyse how emotions, limited foresight, and social learning impact decision-making in simulated situations in order to understand how human beings operate in real-life strategic situations.[39]

c) Areas of Application

Michael Gilligan, Eric Mvukiyehe, and Cyrus Samii use a quasi-experimental design to test the effectiveness of ex-combatant re-integration in Burundi. In their research, instead of randomly assigning participants to treatment and control groups respectively, they exploit bureaucratic failures in the delivery of the re-integration benefits and halts in service delivery to measure programme effects. For future research in this area, the authors of the study advocate for more randomization and within-programme experiments.[40] However, not everyone is convinced of the research merits of randomized designs. In the field of organized crime, Michael Maltz argues there are legal, administrative, and ethical difficulties of using experimental design, given that the control group would be free of law enforcement efforts creating the potential for violence and crime to continue.[41]

d) Advantages

One of the main advantages of experiments is the level of precision and reliability of the findings, as compared to other research methods such as surveys. Experiments allow

researchers to demonstrate in an easily understandable and replicable way how they came to their assessment. In addition, randomization enhances experimental assessments' external validity (i.e. findings are generalizable and broader inferences for larger populations can be made). Stepped-wedge designs, which consist of multiple measurement periods, allow for prediction of the long-term impact of programmes, while avoiding some of the ethical arguments against RCTs. Behavioural games allow researchers to test hypotheses that are difficult to verify in real-life scenarios. In addition, behavioural games facilitate the analysis of decision-making processes by assuming that humans are not always rational and self-interested actors, but emotional actors that operate in a social context.

e) Drawbacks

The use of experiments in security intervention assessments is limited by perceived – and in some cases, real – ethical concerns and logistical considerations. Under the assumption that treatment is actually beneficial, it might be perceived as unethical to deny a population access to treatment in order to establish a control group (e.g. withholding reintegration benefits to some DDR participants) for a long period of time. Moreover, the creation of experimental conditions is logistically complex and often cost-intensive. In particular randomized controlled trials conducted on a large scale require large samples and thus sufficient financial resources, expertise as well as political will to push for and invest in the use of experiments on a larger scale. Given the difficulties of reproducing real-life conditions and the concerns related to research ethics in randomized experiments, quasi-experimental designs and stepped-wedge designs may be important alternate tools.

5. INDICATORS, INDICES, AND BENCHMARKING

a) Description

Indicators, indices, and benchmarking are frequently used to measure progress and to demonstrate the impact of security interventions. An indicator is a "quantitative or qualitative variable that provides a simple and reliable means to measure achievement, to reflect the changes connected to an intervention, or to help assess the performance of a program."[42] Indicators can be designed to reflect both the positive and negative impact of a programme. An index is an "accumulation of scores from a variety of individual indicators that rank-orders specific observations in order to represent a more general concept."[43] Benchmarking is a method of using reference points (i.e. benchmarks) to assess the performance of an intervention.[44] Multiple identity indicators are used to "measure progress toward or regression away from [the chosen] benchmarks."[45]

b) Types

- **Indices/Composite indicators** are used to "measure multi-dimensional concepts, which cannot be captured by a single indicator."[46] They consist of several – often weighted – individual indicators. Examples of the use of composite indicators include the Human Development Index, which consists of several measures, such as life expectancy, education, and per capita income; and the Freedom House Index, which includes several measures of civil liberties and political rights.
- **Proxy indicators** are used where direct measurement is not possible. Proxy indicators offer a way to measure more abstract concepts, such as trust and political integration.
- **Performance indicators** are variables that "allow the verification of changes in the intervention… and measure to what extend objectives are being achieved."[47]
- **Impact indicators** are "variables that allow the assessment of positive and negative, primary and secondary long-term results produced by an intervention. These results can be produced directly or indirectly, and can be intended or unintended."[48]

c) Areas of Application

Indicators and indices play an important role in measuring the effectiveness of security interventions. In the justice sector, the 2003 report of the Vera Institute of Justice examined performance indicators across the justice sector to measure progress toward justice and security.[49] In drug policy research, Robert Muggah proposes a set of indicators to measure the effectiveness of a new drug policy, which are focused on safety, citizen security, and improvements in public health instead of easily quantifiable data enforcement outputs such as the number of arrests or seizures.[50] In 2011, the United Nations Rule of Law Indicators were developed to measure the performance, transparency, and accountability of rule of law institutions and serve as a "diagnostic tool to refine interventions to address the most pressing problems."[51]

d) Advantages and Drawbacks

Indicators enable researchers to measure the progress towards a predetermined goal and thus, the impact of an intervention. Proxy indicators enable an approximate measurement in cases where the phenomenon is not directly measurable. However, if not thoughtfully chosen, indicators may not accurately measure the impact of the intervention at hand. While composite indicators allow for a big picture, they are only as good as their components and the weighting formula behind them, and without a good understanding of the multiple factors at hand, can be used to draw simplistic, non-robust policy conclu-

sions.[52] Such pitfalls are not inevitable and can be overcome by understanding the individual indicators used in the composite, such as the various measures of freedom in the Freedom House Index.

6. SELECTED QUALITATIVE APPROACHES

a) Case Studies

Case study methods "involve systematically gathering enough information about a particular person or ... event to understand how it operates or functions"[53] for the purpose of gaining a deeper understanding about the particularities of the respective case and potentially deriving lessons learned for other cases. Case studies can focus on different units of analysis and can incorporate a number of data gathering measures, such as in-depth interviews, content analysis, and participant observation.[54]

Case studies can take the form of within-case studies, which are an in-depth exploration of sub-units of a single case, and cross-case studies, which allow for comparative studies, highlighting similarities and differences across cases.[55] Case studies can also be divided into sub-types depending on the purpose they are used for: intrinsic case studies seek to provide a better understanding of the intrinsic characteristics of a particular, unique case, and thus have little external value for other cases.[56] Conversely, instrumental case studies serve as an instrument to better develop broader theoretical questions and more general hypotheses that have relevance beyond the case itself.[57] Collective case studies "involve the extensive study of several instrumental cases" to help theorize about a broader context.[58]

In the area of CVE, James Khalil and Martine Zeuthen use Kenya as a case study to examine the USAID Office of Transition Initiatives' CVE pilot project and to draw lessons learned from the Kenyan case study for the broader CVE community.[59] In a 2010 report on organized crime, UNODC employs various regional case studies to highlight the impact of drug trafficking, mineral smuggling, and maritime piracy.[60] While case studies grant in-depth insights into a particular case of relevance, allowing for a better understanding of causal relationships and nuances, subjective decisions by the researcher may raise questions about the objectivity of the results.[61] Moreover, due to context-specific factors, the generalizability of the findings of case studies can be limited.[62]

b) Textual Analysis

Textual analysis "involves the identification and interpretation of a set of verbal or non-verbal signs. The meaning of these signs can be analysed from the perspective of the speaker's intent, the audience's reaction, the historical and cultural context in which the text was created or the

contemporary… context in which the text is experienced today."[63] Textual analysis, which has qualitative and quantitative applications depending on the scale and depth of the approach, is a method researchers use to describe and interpret the characteristics of different types of text.[64] The units of analysis can range from words to semantics, characters, and concepts.[65]

In their research on CVE efforts, Lazar Stankov et al. use a linguistic analysis of texts produced by known terrorist organizations to explore the development of a militant extremist mindset.[66] While textual analysis enables the researcher to understand a particular culture, social group, or phenomenon,[67] the drawback of textual analysis is that the researcher's interpretation is "only one of many possible valid interpretations."[68]

7. TOWARDS A MORE RIGOROUS IMPACT ASSESSMENT

Skilful combination of assessment methods and tools enables more rigorous and effective impact assessment of security interventions. In this section, some of the key practical techniques that can facilitate such approaches are considered: early integration of M&E into programme design; increased time horizons; as well as the use of mixed methods; experimental designs; and new technologies.

a) Early Integration of M&E into Programme Design

Quality impact assessment requires that M&E is viewed as an integral part of programming and that it is integrated into programme design at the conceptualization stage. Unfortunately, impact assessment is often only implemented once programmes are already running and accountability to donors needs to be demonstrated. Budget constraints, lack of resources and expertise, and a lack of political will can prevent the early integration of impact assessment into programme design. Implementing M&E after a programme is already in place has numerous disadvantages, including the failure to establish baselines and gather baseline data against which programmatic progress can later be measured.

Effective impact assessment also needs an adequate strategic framework. The theory of change is a methodology that involves defining long-term goals and subsequently mapping backward short-, medium-, and long-term preconditions, underlying assumptions, and causal relationships that are required to reach this goal.[69] Using a theory of change thus allows programme managers to compare actual outputs to desired impacts and to make more informed decisions about strategy and tactics. However, to include rigorous impact assessment from the conceptualization stage of programmes, long-term impact assessment beyond output evaluations also needs to be prioritized and valued by programmers and donors alike.

b) Lengthen Time Horizons

In order to adequately capture the impact of security interventions, programmers may need to consider longer-term assessment approaches. By employing longitudinal assessments, which study the same population over multiple data-collection periods, programmers will be able to measure impact, not just programmatic outputs.[70] Longitudinal studies avoid some of the biases that plague other research methods (e.g. 'cohort effects', which refer to the effects of shared experiences of a group), but are subject to their own challenges (e.g. 'attrition effects', wherein participants drop out before the end of the study). While drawbacks to this approach must be addressed, the challenges to longer-term assessment are not primarily methodological in nature, but rather political, financial, and bureaucratic. Of particular note are the short-term orientation of mandates and budgetary constraints that prevent shifting assessment timeframes. In the case of UN peacekeeping operations, budgetary reporting cycles are usually one year and since peacekeeping missions leave the country once their mandate ends, it is difficult to fund and conduct longitudinal studies to assess the long-term impact of security interventions. In addition, donors may need to re-calibrate their expectations about effective M&E and be willing to wait longer for comprehensive (and more accurate) results as well as demonstrate a willingness to deal with non-results. Managing donor expectations will reduce the reporting of unreliable data just because donors demand a demonstration of results.

c) Mixed Methods

As each approach to assessment has particular advantages and drawbacks, a combination of methods can help overcome constraints and achieve a more comprehensive picture of a security intervention's impact.[71] Mixing various methods allows researchers to approach problems from multiple ways with the hope that, despite the drawbacks associated with any particular approach, the use of more than one, will improve the robustness of the findings. Method triangulation can be employed both simultaneously or sequentially, by using different metrics and applying multiple research methods (e.g. quantitative and qualitative approaches). As the methods employed become more numerous and potentially complex, however, practitioners need to become more conversant in methods and understand the implications of the methods and metrics used to assess their programmes.

d) Experimental Design

Experiments bear great potential for impact assessment of security interventions. While potential obstacles can arise when there is resistance to experiments, it is often possible to mitigate such concerns with thoughtful design. For example, stepped-wedge designs seem to be particularly promising for the field of security interventions. In these designs, participants are divided into treatment and control groups. However, instead of denying the control group treatment entirely, the control group receives treatment at a different point in time.

In many ways, stepped-wedge studies mimic the logistical realities of rolling out interventions, whereby it is rare that all groups will be exposed to the intervention simultaneously. These designs open new possibilities in experimental designs as they can help to address ethical concerns about denying beneficiaries potentially life-saving or –enhancing treatments.

e) New Technologies

Information and communication technologies in both data collection and data analysis have put more comprehensive, accurate, and less expensive measurement within reach. In particular in non-conducive security environments, these technologies have potential utility in improving access to beneficiaries, and thus in "their ability to produce large data sets, which can be combined with other tools for greater analytical capacities and insights."[72]

Cellular technology and cloud computing have made it easier to poll people in dangerous areas and analyse their political preferences.[73] For example, in the rule of law field, the low cost and pervasiveness of mobile technologies is facilitating perception-survey based methodologies for rapid assessment of citizen attitudes to different justice needs and providers.[74] Moreover, digital technologies are also transforming the sources of data available to researchers and funders, and the ways that communities impacted by security interventions can participate in monitoring and accountability. In South America, for example, the availability of digital cameras and social media is transforming approaches to monitoring policing and complex security interventions such as the pacification programmes in Brazil's favelas.[75]

Information and communication technologies can also serve as a tool to facilitate data collection and analysis, given their ability to collect and analyse large data sets and to potentially reach even rural areas: Even on phones without WiFi or GPS, SMS and phone calls can be used as a tool to manage data collection, as in the case of the FrontlineSMS software that was employed in Search For Common Ground's Promoting Inclusive and Participatory Elections (PIPE) in the Democratic Republic of the Congo. FrontlineSMS has been used to "market peacebuilding radio programmes through SMS blasts."[76] Real-time automated data aggregation and data analysis with the help of software makes these new technologies a cost- and time-efficient tool. Yet given the privacy concerns associated with such information and communication technologies it is important to place these technologies duly under a "strict political and legal framework [that addresses] control and confidentiality concerns."[77]

All of these attempts to make impact assessment in contemporary security interventions more rigorous are, however, to no avail if the findings are not worked back into programmatic planning. Consequently, attempts need to be made not only to systematically include rigorous impact assessments in security interventions, but also to modify programmes where needed and to draw lessons for future project design.

8. CONCLUSION

The analysis of evolving approaches to impact assessment suggests that assessing the effectiveness of contemporary security interventions remains, despite innovations and technological advances, a challenging endeavour. Despite, and in the face of, these challenges, the practitioner community needs to continue find ways to improve its approach to impact assessment to ensure that programming is effective and efficient. An empirics-led iterative programming cycle, based on a sound theory of change and research design, is required to conduct high-quality impact assessment that measures not easily quantifiable outputs, but metrics that are actually informative regarding the effectiveness of security interventions. Inadequate impact assessments can not only lead to a misinterpretation of programmatic impacts, but may also allow ineffective and even harmful programming, which can have detrimental effects to beneficiaries as well as society at large, to continue.

As laid out here, there are numerous possibilities for improving the impact assessment of security interventions. Introducing randomization into M&E can – when done well – address ethical concerns and ensure robust assessments. The use of information and communication technologies in both data collection and data analysis has put more accurate and cost-effective impact assessment within reach. While this paper has acknowledged that these innovations are not a cure to all problems, they do provide potential ways to improve upon the current state of assessment in this area.

What does effective impact assessment require beyond methodological considerations? First of all, a cultural change is necessary: practitioners need to become more conversant in different research methods and tools. This will enable them to measure impact beyond easily quantifiable measures and to combine methods where necessary. It will also make practitioners better situated for interpreting results, incorporating impact assessment into planning, and choosing the right assessment experts and contractors. Likewise, impact assessment needs to be embraced not just as a necessary box to tick, but as an integral part of programme planning; it should be incorporated at the conceptualization stage of intervention planning. Third, cultural sensitivity and contextual knowledge is required to design effective assessment exercises for security interventions to ensure that biased sampling does not cloud the evaluation of intervention impacts. Fourth, a careful recalibration of expectations and time frames, and thus budgets, is necessary at the programmatic, bureaucratic, and donor level. Such a shift would be encouraged if donors enhance their M&E requirements, provide line-item funding for assessment, and extend funding cycles to account for longitudinal assessment timeframes. By moving from short-term political priorities to long-term studies, practitioners will be able to shift from measuring outputs to measuring impact, and thus get a clearer picture of the effects of security interventions.

Lastly, and on a related note, the professionals conceptualizing and executing security intervention in the areas of rule of law, drug control policy, organized crime, and CVE need to acknowledge that better approaches to assessment may demonstrate that some security interventions have lacklustre impacts. Rather than adopt a bureaucratic survival approach that views such results as damaging, the community needs to shift its mindset to embrace rigorous measurement as the first step to identifying lessons learned, tailoring programming, and improving the impacts of security interventions on the lives of vulnerable peoples. While it will be necessary to address and navigate the challenges identified herein in the future, the assessment tools and methods discussed in this paper indicate that a more accurate and cost-effective measurement is within reach.

ENDNOTES

[1] In this context, the term "security interventions" is used to describe the conceptualization and execution of programmes in the areas of the rule of law, drug control policy, organized crime, and countering violent extremism (CVE).

[2] United Nations General Assembly, "Transforming our world: the 2030 Agenda for Sustainable Development", A/70/L.1.

[3] United Nations, "Open Working Group proposal for Sustainable Development Goals," available from https://sustainabledevelopment.un.org/focussdgs.html (accessed 25 June 2015).

[4] United Nations General Assembly, "Transforming our world: the 2030 Agenda for Sustainable Development", A/70/L.1.

[5] Robert Muggah, Katherine Aguirre and Ilona Szabo de Carvalho, "Measurement Matters: Designing New Metrics for a Drug Policy that Works," Strategic Paper 12, Igarapé Institute, January 2015, p. 2.

[6] Julia Anderson, Julia Anderson, Fernando Bouzas, Rashid Dar, Yohsuke Fukamachi, Jiri Jelinek, Sophie Klinger, Amanda Roth, and Felipe Umaña, "Evaluating Success in Tackling Transnational Organized Crime," Capstone Project Report, Columbia University, May 2015, p. 6.

[7] Ana Leao, "Chapter 1: Disarmament Initiatives in Mozambique," in: Institute for Security Studies Africa, Weapons in Mozambique: Reducing Availability and Demand, available from https://issafrica.org/pubs/Monographs/no94/Chap1.pdf (accessed 7 July 2015).

[8] Marie Eloïse Muller, *From Warfare to Welfare: Human Security in a Southern African Context,* (Assen, The Netherlands: Royal Van Gorcum, 2004), p. 31; and Jennifer Perry, "Small Arms and Light Weapons Disarmament Programs: Challenges, Utility, and Lessons Learned," US Defense Threat Reduction Agency, 12 July 2004, p. 10.

[9] Mark Knight and Alpaslan Ozerdeem, "Guns, Camps, and Cash: Disarmament, Demobilization, and Reinsertion of Former Combatants in Transitions from War to Peace," *Journal of Peace Research*, Vol. 41, No. 4, July 2004, pp. 501-502.

[10] Anderson et al., p. 6.

[11] Muggah et al., p. 4.

[12] For example, see Martin Gramatikov, Kavita Heijstek-Ziemann, Roger El Khoury, Gediminas Motiejunas, Sam Muller, and David Osborne, "Family, Justice and Fairness in Yemen: The Impact of Family Problems on Yemeni Women," Hiil publication, 2014.

[13] Susan Berg-Seligson, Diana Orces, Georgina Pizzolitto, Mitchell Seligson and Carole Wilson, "Impact evaluations of UNAID's Community-Based Crime and Violence Prevention approach in Central America," Latin America Public Opinion Project, 2014.

[14] Anthony A Braga and David L. Weisburd, "The Effects of 'Pulling Levers' Focused Deterrence Strategies on Crime," *Campbell Systematic Reviews*, Vol. 6, March 2012, available from http://www.campbellcollaboration.org/lib/project/96/ (accessed 30 July 2015).

[15] Not every qualitative and quantitative method available will be detailed in this section, but rather the focus will be on the most relevant research methodologies, metrics, and tools for assessing contemporary security interventions. As with all M&E, the specific choice of methods depends on the research questions, resources, context, and programme size at hand.

[16] Scott van der Stoep and Deirdre Johnston, Research Methods for *Everyday Life – Blending Qualitative and Quantitative Approaches*, (San Francisco: John Wiley and Sons Publication, 2009), p. 37.

[17] University of Surrey, "Questionnaires", available from http://libweb.surrey.ac.uk/library/skills/Introduction%20to%20Research%20and%20Managing%20Information%20Leicester/page_48.htm (accessed 15 July 2015).

[18] Bruce Berg, *Qualitative Research Methods for Social Sciences*, Fourth edition, (Needham Heights: Allyn & Bacon, 2001), p. 111.

[19] Ibid., p. 70.

[20] HIIL, "The Needs of the Malians for Justice: Towards More Fairness," 2014, available from http://www.hiil.org/data/sitemanagement/media/Mali%20Report_HiiL_Final%20English_low_resolution.pdf (accessed 12 July 2015).

[21] Christopher Blattman and Jeannie Annan, "Child Combatants in Northern Uganda: Reintegration Myths and Realities," in Robert Muggah, ed., *Security and Post-Conflict Reconstruction*, (New York City, Routledge, 2008), p. 103–126.

[22] Macartan Humphreys and Jeremy Weinstein, "Demobilization and Reintegration," *Journal of Conflict Resolution*, Vol. 51, No. 4, August 2007, p. 531-567.

[23] Vanessa Corlazzoli, *Baseline Report: Countering and Preventing Radicalisation In Indonesian Pesantrens*, (Washington, DC: Search for Common Ground, 2011).

[24] Martin Gramatikov, et al.

[25] Scott van der Stoep and Deirdre Johnston, *Research Methods for Everyday Life – Blending Qualitative and Quantitative Approaches*, (San Francisco: John Wiley and Sons Publication, 2009), p. 37.

[26] Bruce Berg, *Qualitative research methods for social sciences*, Fourth edition, (Needham Heights, Allyn & Bacon, 2001), p. 82.

[27] Ibid., p. 74.

[28] For example, being denied access to programmes. van der Stoep and Johnston, p. 185.

[29] Berg, p. 85.

[30] Ibid., p. 80.

[31] Ibid., p. 27.

[32] Ibid., p. 31.

[33] To determine that a causal relationship exists, a researcher will try to find evidence that a factor X produces Y and test whether in the absence of X, there is no Y. See van der Stoep and Johnston, p. 106.

[34] Ibid., p. 112.

[35] Ibid., p. 131.

[36] The objective of RCTs is to reduce the influence of confounding variables - those variables that can lead to incorrect interpretations of the relationship between the dependent and the independent variables - through randomization. Ibid., p. 120.

[37] Ibid., p. 147.

[38] In comparison to repeated measures designs, in which subjects are exposed to all treatments.

[39] Colin Camerer, *Behavioural Game Theory: Experiments in Strategic Interaction*, (Princeton: Princeton University Press, 2013).

[40] Michael Gilligan, Erik Mvukiyehe and Cyrus Samii, "Reintegrating Rebels into Civilian Life: Quasi-experimental Evidence from Burundi," *Journal of Conflict Resolution*, Vol. 57, No. 4, 2012, p. 21.

[41] Michael D. Maltz, *Measuring the Effectiveness of Organized Crime Control Efforts*, (Chicago: University of Illinois at Chicago, Office of International Criminal Justice, 1990), p. 15.

[42] Organisation for Economic Co-operation and Development (OECD), "Glossary of Key Terms in Evaluation and Results Based Management," 2002, available from http://www.oecd.org/development/peer-reviews/2754804.pdf (accessed 2 September 2015), p. 25.

[43] Vanessa Corlazzoli and Jonathan White, "Measuring the Un-measurable – Solutions to Measurement Challenges in Fragile and Conflict Affected Environments," Search for Common Ground/UK Department for International Development, March 2013, p. 15.

[44] OECD, "Glossary of Key Terms in Evaluation and Results Based Management," p. 18.

[45] United Nations, "Monitoring Peace Consolidation – United Nations Practitioners' Guide to Benchmarking," 2010, available from http://www.un.org/en/peacebuilding/pbso/pdf/monitoring_peace_consolidation.pdf (accessed 2 September 2015).

[46] OECD, "The OECD-JRC Handbook on Practices for Developing Composite Indicators," paper presented at the OECD Committee on Statistics, 7-8 June 2004, OECD, Paris, available from http://stats.oecd.org/glossary/detail.asp?ID=6278 (accessed 30 July 2015).

[47] Ibid.

[48] Ibid.

[49] Vera Institute of Justice, "Measuring Progress Towards Safety and Justice: A Global Guide to the Design of Performance Indicators Across the Justice Sector," November 2003.

[50] Muggah et al.

[51] United Nations, "Rule of Law Indicators: Implementation Guide and Project Tools," 2011, available from http://www.un.org/en/events/peacekeepersday/2011/publications/un_rule_of_law_indicators.pdf (accessed 12 July 2015).

[52] European Commission Composite Indicators Research Group, "What is a Composite Indicator?" 18 December 2014, available from https://composite-indicators.jrc.ec.europa.eu/?q=content/what-composite-indicator (accessed 2 September 2015).

[53] Berg, p. 225.

[54] Ibid.

[55] Ibid.

[56] Ibid, p. 229.

[57] Ibid.

[58] Ibid.

[59] James Khalil and M. Zeuthen, "A Case Study of Counter Violent Extremism (CVE) Programming: Lessons from OTI's Kenya Transition Initiative," *Stability: International Journal of Security and Development*, Vol. 3, No. 1, (2014), p. 31.

[60] UNODC, "Crime and Instability: Case Studies of Transnational Threats," February 2010.

[61] Berg, p. 231.

[62] van der Stoep and Johnston, p. 210.

[63] Ibid., p. 211.

[64] Sage Research Methods, "What is Textual Analysis?" available from http://srmo.sagepub.com/view/textual-analysis/n1.xml (accessed 12 July 2015).

[65] Berg, p. 247.

[66] Lazar Stankov Derrick Higgins, Gerard Saucier, and Goran Knezvic, "Contemporary Militant Extremism: A Linguistic Approach to Scale Development," *Psychological Assessment* , Vol. 22, No. 2, (2010), p. 246–258.

[67] van der Stoep and Johnston, p. 213.

[68] Ibid.

[69] Center for Theory of Change, "What is Theory of Change?" available from http://www.theoryofchange.org/what-is-theory-of-change/ (accessed 15 July 2015).

[70] van der Stoep and Johnston, p. 39.

[71] Berg, p. 5.

[72] Corlazzoli and White, p. 22.

[73] For example, see Craig Charney, "Here, There, and Everywhere: The Cell Phone at the Bottom of the Pyramid," December 5, 2009, available from http://www.charneyresearch.com/resources/here-there-and-everywhere-the-cell-phone-at-the-bottom-of-the-pyramid/ (accessed 30 July 2015).

[74] See for example The Hague Institute for the Internationalisation of Law, "Les besoins des Maliens en Matière de Justice: Vers Plus d'Équité," available from http://www.hiil.org/data/sitemanagement/media/HiiL_Mali_Report_lores.pdf (accessed 2 September 2015).

[75] David Bruce and Sean Tait, "A 'Third Umpire' for Policing in South Africa: Applying Body Cameras in Western Cape," available from http://en.igarape.org.br/a-third-umpire-for-policing-in-south-africa/ (accessed 2 September 2015).

[76] Corlazzoli and White, p. 22.

[77] Ibid.